PREDICTION DU PRIX MEDIAN DES MAISONS OCCUPEES PAR LEURS PROPRIETAIRES

BOSTON HOUSE PRICE

NIANG Mamadou

Master II Traitement de l'Information et Data science en entreprise

Année scolaire 2021-2022

I- INTRODUCTION

Aux Etats-Unis comme partout dans le monde le prix de l'immobilier n'a cessé d'être dépendant des différentes caractéristiques du secteur où il se trouve. Même si cela est accepté à l'unanimité, force est de reconnaître qu'une étude statistique rigoureuse sur le sujet n'aura qu'un fort plus dans la consolidation de cette vérité. Ainsi c'est dans ce cadre que s'inscrit notre étude.

En effet dans le présent document nous essaierons de voir les différents facteurs qui peuvent influer sur le prix d'une maison et de quelle façon. Pour mener à bien l'étude nous utilisons une base de données, tirée du Boston Standard Metropolitan Statistical Area (SMSA) en 1970, où sont répertoriées des observations concernant les différentes villes de la SMSA Boston. Aux Etats-Unis, une zone statistique métropolitaine (SMSA) est une région géographique avec une densité de population relativement élevée en son cœur et des liens économiques étroits dans toute la zone.

La problématique clé de l'étude restera de trouver quelles variables influencent la valeur médiane des maisons occupées par leurs propriétaires.

Pour répondre à la question, le document sera scindé en 4 parties. En premier lieu, on passe à la prise en main de la base de données pour ensuite attaquer la démarche méthodologique suivie durant toute l'étude. Avant de conclure, on s'attèlera à la modélisation proprement dite ainsi qu'à l'interprétation des résultats trouvés.

II- EXPLORATION DE LA BASE DE DONNEES

Ici dans cette section on commencera par une prise en main de nos données, chose qui est indispensable pour une étude.

La base de données qu'on utilisera tout au long de cette étude provient du site de la MSA Boston (Zone Statistique Métropolitaine).

Elle est constituée de 506 observations auxquelles sont associées 14 variables. Chaque observation représente une ville de la MSA Boston. Ci-dessus l'entête de la base de données.

Figure 1 : En-tête de la base de données

	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	MEDV
1	0.00632	18.0	2.31	0	0.5380	6.575	65.2	4.0900	1	296	15.3	396.90	4.98	24.0
2	0.02731	0.0	7.07	0	0.4690	6.421	78.9	4.9671	2	242	17.8	396.90	9.14	21.6
3	0.02729	0.0	7.07	0	0.4690	7.185	61.1	4.9671	2	242	17.8	392.83	4.03	34.7
4	0.03237	0.0	2.18	0	0.4580	6.998	45.8	6.0622	3	222	18.7	394.63	2.94	33.4
5	0.06905	0.0	2.18	0	0.4580	7.147	54.2	6.0622	3	222	18.7	396.90	5.33	36.2

Par défaut de ne pouvoir afficher toutes les observations, on essaiera de voir quand même s'il y'a une présence potentielle de valeurs manquantes ou pas.

Figure 2 : Informations sur les variables

variable	q_zeros	p_zeros	q_na	p_na	q_inf	p_inf	type	unique
CRIM	0	0.00	0	0	0	0	numeric	504
ZN	372	73.52	0	0	0	0	numeric	26
INDUS	0	0.00	0	0	0	0	numeric	76
CHAS	471	93.08	0	0	0	0	factor	2
NOX	0	0.00	0	0	0	0	numeric	81
RM	0	0.00	0	0	0	0	numeric	446
AGE	0	0.00	0	0	0	0	numeric	356
DIS	0	0.00	0	0	0	0	numeric	412
RAD	0	0.00	0	0	0	0	numeric	9
TAX	0	0.00	0	0	0	0	numeric	66
PTRATIO	0	0.00	0	0	0	0	numeric	46
B	0	0.00	0	0	0	0	numeric	357
LSTAT	0	0.00	0	0	0	0	numeric	455
MEDV	0	0.00	0	0	0	0	numeric	229

Ce tableau nous donne les informations suivantes :

q_zeros : nombre de zéros de la variable ;

p_zeros : pourcentage de zéros de la variable

; **q_na** : nombre de valeurs manquantes ;

p_na : pourcentage de valeurs manquantes ;

q_inf : nombre de valeurs infinies ; **p_inf** :

pourcentage de valeurs infinies ; **Unique** :

nombre de valeurs uniques

Ce tableau nous informe clairement de l'absence de toute valeurs manques dans notre base de données.

En regardant la colonne type, on voit bien que seule « CHAS » est de type qualitatif, sinon toutes les autres sont quantitatives.

Après on regardera quelques statistiques de ces variables (min, moyenne, max…).

Figure 3 : Statistiques des variables quantitatives

CRIM	ZN	INDUS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	MEDV
Min. : 0.00632	Min. : 0.00	Min. : 0.46	Min. :0.3850	Min. :3.561	Min. : 2.90	Min. : 1.130	Min. : 1.000	Min. :187.0	Min. :12.60	Min. : 0.32	Min. : 1.73	Min. : 5.00
1st Qu.: 0.08205	1st Qu.: 0.00	1st Qu.: 5.19	1st Qu.:0.4490	1st Qu.:5.886	1st Qu.: 45.02	1st Qu.: 2.100	1st Qu.: 4.000	1st Qu.:279.0	1st Qu.:17.40	1st Qu.:375.38	1st Qu.: 6.95	1st Qu.:17.02
Median : 0.25651	Median : 0.00	Median : 9.69	Median :0.5380	Median :6.208	Median : 77.50	Median : 3.207	Median : 5.000	Median :330.0	Median :19.05	Median :391.44	Median :11.36	Median :21.20
Mean : 3.61352	Mean : 11.36	Mean :11.14	Mean :0.5547	Mean :6.285	Mean : 68.57	Mean : 3.795	Mean : 9.549	Mean :408.2	Mean :18.46	Mean :356.67	Mean :12.65	Mean :22.53
3rd Qu.: 3.67708	3rd Qu.: 12.50	3rd Qu.:18.10	3rd Qu.:0.6240	3rd Qu.:6.623	3rd Qu.: 94.08	3rd Qu.: 5.188	3rd Qu.:24.000	3rd Qu.:666.0	3rd Qu.:20.20	3rd Qu.:396.23	3rd Qu.:16.95	3rd Qu.:25.00
Max. :88.97620	Max. :100.00	Max. :27.74	Max. :0.8710	Max. :8.780	Max. :100.00	Max. :12.127	Max. :24.000	Max. :711.0	Max. :22.00	Max. :396.90	Max. :37.97	Max. :50.00

L'objectif étant de modéliser la variable **MEDV**, il est donc nécessaire qu'on regarde les relations potentielles qui puissent exister entre les différentes variables de la base et cette variable en question.

Pour les variables qualitatives, on s'intéresse plus aux coefficients de corrélation de Pearson. On trouve ces corrélations dans le tableau ci-dessous :

Figure 4 : Matrice des corrélations

	CRIM	ZN	INDUS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	MEDV
CRIM	1.0000000	-0.2004692	0.4065834	0.4209717	-0.2192467	0.3527343	-0.3796701	0.6255051	0.5827643	0.2899456	-0.3850639	0.4556215	-0.3883046
ZN	-0.2004692	1.0000000	-0.5338282	-0.5166037	0.3119906	-0.5695373	0.6644082	-0.3119478	-0.3145633	-0.3916785	0.1755203	-0.4129946	0.3604453
INDUS	0.4065834	-0.5338282	1.0000000	0.7636514	-0.3916759	0.6447785	-0.7080270	0.5951293	0.7207602	0.3832476	-0.3569765	0.6037997	-0.4837252
NOX	0.4209717	-0.5166037	0.7636514	1.0000000	-0.3021882	0.7314701	-0.7692301	0.6114406	0.6680232	0.1889327	-0.3800506	0.5908789	-0.4273208
RM	-0.2192467	0.3119906	-0.3916759	-0.3021882	1.0000000	-0.2402649	0.2052462	-0.2098467	-0.2920478	-0.3555015	0.1280686	-0.6138083	0.6953599
AGE	0.3527343	-0.5695373	0.6447785	0.7314701	-0.2402649	1.0000000	-0.7478805	0.4560225	0.5064556	0.2615150	-0.2735340	0.6023385	-0.3769546
DIS	-0.3796701	0.6644082	-0.7080270	-0.7692301	0.2052462	-0.7478805	1.0000000	-0.4945879	-0.5344316	-0.2324705	0.2915117	-0.4969958	0.2499287
RAD	0.6255051	-0.3119478	0.5951293	0.6114406	-0.2098467	0.4560225	-0.4945879	1.0000000	0.9102282	0.4647412	-0.4444128	0.4886763	-0.3816262
TAX	0.5827643	-0.3145633	0.7207602	0.6680232	-0.2920478	0.5064556	-0.5344316	0.9102282	1.0000000	0.4608530	-0.4418080	0.5439934	-0.4685359
PTRATIO	0.2899456	-0.3916785	0.3832476	0.1889327	-0.3555015	0.2615150	-0.2324705	0.4647412	0.4608530	1.0000000	-0.1773833	0.3740443	-0.5077867
B	-0.3850639	0.1755203	-0.3569765	-0.3800506	0.1280686	-0.2735340	0.2915117	-0.4444128	-0.4418080	-0.1773833	1.0000000	-0.3660869	0.3334608
LSTAT	0.4556215	-0.4129946	0.6037997	0.5908789	-0.6138083	0.6023385	-0.4969958	0.4886763	0.5439934	0.3740443	-0.3660869	1.0000000	-0.7376627
MEDV	-0.3883046	0.3604453	-0.4837252	-0.4273208	0.6953599	-0.3769546	0.2499287	-0.3816262	-0.4685359	-0.5077867	0.3334608	-0.7376627	1.0000000

En analysant la dernière ligne ou a dernière colonne de ce tableau, on remarque bien que toutes les vari ables quantitatives sont corrélées avec la variable MEDV. On prendra donc toutes ces variables dans notre modèle et voir si elles sont significatives ou non. En regardant cette matrice de corrélations, on peut aussi soupçonner d'autres problèmes comme celui de la multicollinéarité. Mais ces problèmes seront abordés dans les parties qui suivent.

La variable CHAS étant la seule variable qualitative de la base et ne comprenant que 2 modalités, on va donc premièrement voir si la moyenne de MEDV diffère en fonction de la valeur prise par CHAS (0 ou 1).

	CHAS = 0	CHAS = 1
Moyenne de MEDV	22.09384	28.44000

On voit que les 2 moyennes sont différentes même si l'écart ne semble pas être grand. Ce qui signifie que cette variable influe sur MEDV. Donc on la tiendra compte dans notre modèle.

III- DEMARCHE METHODOLOGIQUE

Comme évoqué ci-dessus, l'objectif principal de cette étude est de dresser un modèle capable de prédire le prix médian d'une maison et cela à partir de différentes variables explicatives.

Notre variable cible (Prix médian) est de type quantitatif continu. Quant aux variables explicatives, elles sont toutes quantitatives à l'exception d'une seule (CHAS).

Dans la littérature, le modèle classique de régression utilisé pour expliquer une variable quantitative continue par un ensemble d'autres variables est la régression linéaire multiple. D'où notre motivation pour ce modèle.

1- Spécification du modèle :

En fonction du type des données on distingue 2 spécifications possibles d'un modèle de régression linéaire multiples. Les modèles en séries temporelles (données indicées dans le temps) et les modèles en coupes instantanées (données observées sur un ensemble d'individus). Ici dans cette étude on utilisera la seconde spécification.

Ainsi on aura :

$$Y_i = \theta X_i + U_i \quad \text{pour} \quad i = 1 \text{ à } N$$

Yi : variable à expliquer pour l'individu i
Xi : matrice des variables explicatives pour l'individu i
θ: vecteur des paramètres du modèle à estimer
N : nombre d'observations
Ui : erreur de spécification du modèle.

Afin d'estimer de manière rigoureuse et approximer au mieux la réalité, la validité de certaines hypothèses sont cruciales :

H1 : le modèle est linéaire en X_i
H2 : X_i est non aléatoire, ses valeurs sont observées sans erreurs
H3 : En moyenne le modèle est bien spécifié ($E(U_i) = 0$)
H4 : La variance de l'erreur est constante ($E(U_i^2) = \sigma^2$
H5 : Les erreurs sont non corrélées ($E(U_i U_{i'}) = 0$) si i ≠ i ′)

2- Estimation du modèle :

On dispose de plusieurs techniques d'estimation du modèle précédemment posé.

Mais si les hypothèses ci-dessus sont vérifiées, la méthode d'estimation la plus utilisée est celle des moindres carrés ordinaires (MCO), qui consiste à minimiser la somme des carrés des résidus.

Pour des soucis d'espace on ne la développera pas ici. Mais la formule close de l'estimateur issu de cette méthode est la suivante :

$$\hat{\theta} = (X'X)^{-1}X'Y$$

L'estimateur est sans biais, c'est-à-dire que sa moyenne donne la vraie valeur du paramètre estimé θ.

Sa variance est très importante dans la mesure où elle nous permet de faire l'inférence statistique et de construire des intervalles de confiance. Sa formule est donnée par :

$$Var(\hat{\theta}) = \sigma^2(X'X)^{-1}$$

3- Vérification des hypothèses :

La violation de certaines hypothèses conduirait à des estimations biaisées et par conséquent à des conclusions fallacieuses. Après estimation il est donc important de bien vérifier si elles sont valides.

Ces hypothèses sont entre autres :

- ***La normalité des résidus :*** Pour la vérifier on utilise le QQ-plot (graphique) ou le test de Kolmogorov-Smirnov (KS). En général cette hypothèse est toujours vérifiée quand on dispose d'un grand nombre d'observations.
- ***L'Homoscédasticité des résidus :*** Pour la vérifier, on utilise le test Breusch-Pagan ou celui de White. Si elle est violée, l'estimateur MCO reste tout de même sans biais. Mais pour faire de l'inférence statistique, on est obligé d'utiliser une autre matrice variance-covariance robuste à l'hétéroscédasticité (par exemple la matrice variance covariance de White).
- ***Autocorrélation des résidus :*** Pour la vérifier, on utilise le test de Breusch-Godfrey (BG test) qui est une généralisation du test de Durbin-Watson et le test de Box Pierce. Avec les données en coupes instantanées on peut souvent faire fi à cette hypothèse.
- ***Multicollinéarité :*** apparaît si 2 ou plusieurs variables explicatives semblent corrélées entre elles. Ce phénomène pose un problème d'inversibilité, qui est pourtant nécessaire pour l'estimateur MCO. On la détecte à l'aide du VIF (Facteur d'Inflation de la Variance). On supprime la variable qui a le plus grand VIF s'il est supérieur à 5.

4- Qualité du modèle et différents tests :

Le critère qui est fondamentalement utilisé pour mesurer la qualité d'un modèle de régression linéaire multiple est le coefficient de détermination. Ce coefficient mesure à quel point le modèle ajuste nos données. Ainsi plus on a un bon modèle, plus ce coefficient augmente. Ce coefficient se calcule par la formule suivante :

$$R^2 = \frac{SCE}{SCT}$$

Mise à part ce coefficient on peut aussi de façon individuelle ou globale tester la significativité des variables explicatives utilisées dans notre modélisation.

Pour le tester la significativité individuelle on utilise le test Student et quant à la significativité globale on fait appel au test de Fisher (généralisation du test de Student).

Souvent il n'est pas nécessaire de garder certaines variables dans notre modèle. Ainsi pour choisir ceux à garder, on fait appel aux stratégies de sélection de variables. Ces stratégies se basent toutes sur des comparaisons de critères issues des différents modèles. Par conséquent on choisit les variables dont la combinaison nous donne la meilleure performance.

Ici dans notre étude on utilise la stratégie *backward* pour sélectionner nos variables.

IV- MODELISATION

1- Estimation du modèle :

Dans un premier temps, nous allons tester la possibilité d'un modèle complet qui prendra en compte toutes les variables. Après ajustement, on obtient les résultats ci-dessous :

Figure 5 : Estimation du modèle saturé

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.646e+01  5.103e+00   7.144 3.28e-12 ***
CRIM        -1.080e-01  3.286e-02  -3.287 0.001087 **
ZN           4.642e-02  1.373e-02   3.382 0.000778 ***
INDUS        2.056e-02  6.150e-02   0.334 0.738288
CHAS         2.687e+00  8.616e-01   3.118 0.001925 **
NOX         -1.777e+01  3.820e+00  -4.651 4.25e-06 ***
RM           3.810e+00  4.179e-01   9.116  < 2e-16 ***
AGE          6.922e-04  1.321e-02   0.052 0.958229
DIS         -1.476e+00  1.995e-01  -7.398 6.01e-13 ***
RAD          3.060e-01  6.635e-02   4.613 5.07e-06 ***
TAX         -1.233e-02  3.760e-03  -3.280 0.001112 **
PTRATIO     -9.527e-01  1.308e-01  -7.283 1.31e-12 ***
B            9.312e-03  2.686e-03   3.467 0.000573 ***
LSTAT       -5.248e-01  5.072e-02 -10.347  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.745 on 492 degrees of freedom
Multiple R-squared:  0.7406,    Adjusted R-squared:  0.7338
F-statistic: 108.1 on 13 and 492 DF,  p-value: < 2.2e-16
```

Commentaires de la sortie

Test de significativité globale : ***(Test de Fisher)***

Le résultat du test de significativité globale nous suggère de rejeter l'hypothèse H0 (H0 : tous les coefficients sont nuls sauf la constante) car p-value très proche de 0.

R2 :

La variabilité associée au prix d'un logement à Boston (**MEDV**) est expliquée à **74%** par notre modèle c'est à dire les variables qu'on a considéré dans le modèle.

Test de nullité des paramètres estimés : ***(Test de Student)***

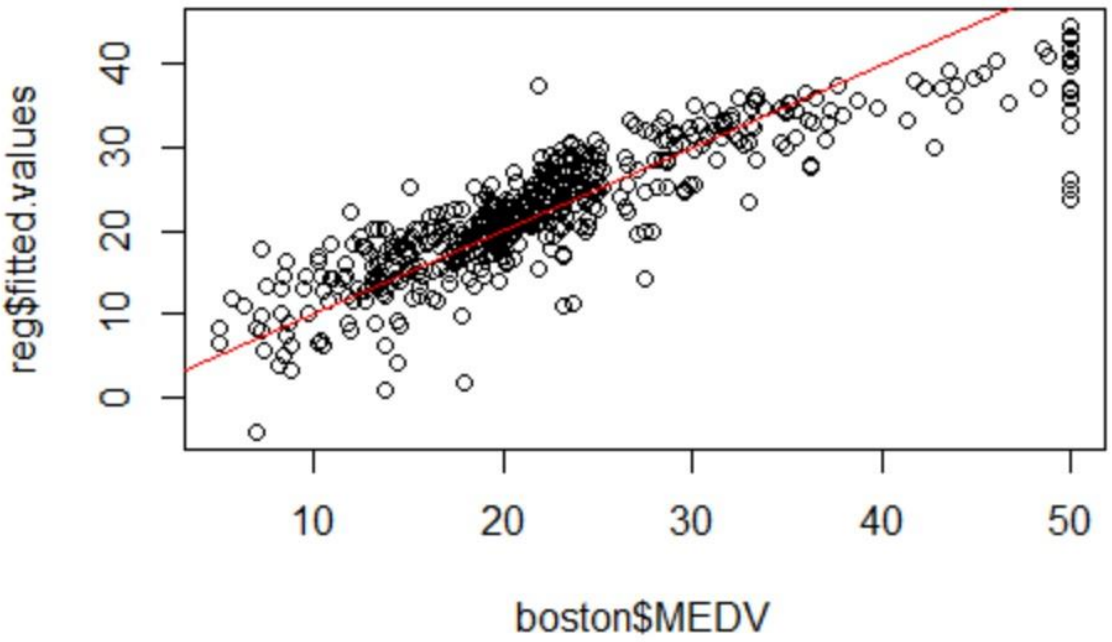
Variables significatives : CRIM, ZN, CHAS, NOX, RM, DIS, RAD, TAX, PTRATIO, B, LSTAT.

Donc ces variables ont une influence sur le prix du logement.

Variables non significatives : INDUS et AGE. Soit ces variables n'influent pas, soit il y'a une forte colinéarité entre les variables explicatives.

Graphique des valeurs prédites par rapport aux valeurs observées

Figure 6 : Graphique des valeurs prédite par rapport aux valeurs observées



On remarque que le nuage de points obtenu est proche de la première bissectrice donc le modèle globalement bon.

2- Sélection de modèle avec critère BIC:

Ci-dessous, on a la sortie de la dernière étape de notre sélection de modèle.

Figure 7 : Sélection de modèle

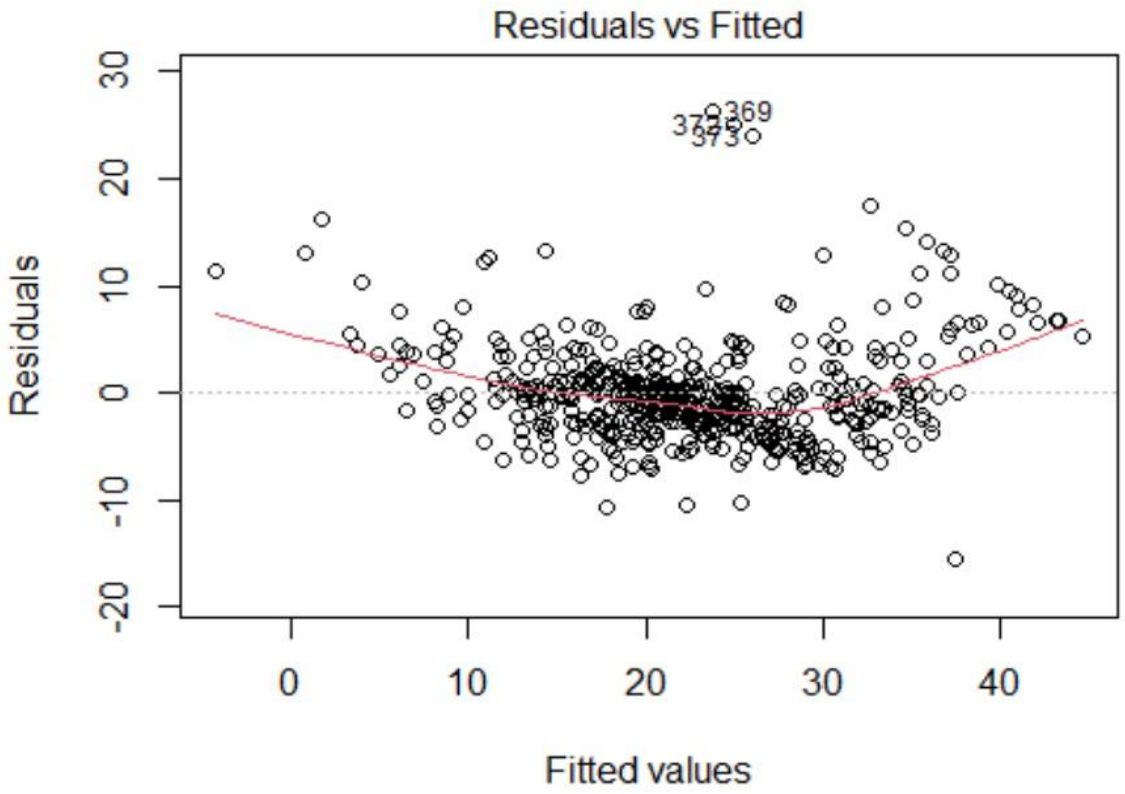
```
Step:  AIC=1636.48
boston$MEDV ~ CRIM + ZN + CHAS + NOX + RM + DIS + RAD + TAX +
    PTRATIO + B + LSTAT

          Df Sum of Sq   RSS    AIC
<none>                  11081 1636.5
- CHAS     1    227.21 11309 1640.5
- CRIM     1    245.37 11327 1641.3
- ZN       1    257.82 11339 1641.9
- B        1    270.82 11352 1642.5
- TAX      1    273.62 11355 1642.6
- RAD      1    500.92 11582 1652.6
- NOX      1    541.91 11623 1654.4
- PTRATIO  1   1206.45 12288 1682.5
- DIS      1   1448.94 12530 1692.4
- RM       1   1963.66 13045 1712.8
- LSTAT    1   2723.48 13805 1741.5
```

Le modèle sélectionné pour expliquer le prix d'un logement est celui qui contient les variables : CRIM, ZN, CHAS, NOX, RM, DIS, RAD, TAX, PTRATIO, B, LSTAT. On retrouve ainsi le modèle qui contient les variables significatives seulement.

Analyse des résidus :

Figure 8 : Graphique des résidus par rapport aux valeurs prédites



On remarque bien que le nuage de points n'est pas bien « équilibré » autour de l'axe des abscisses. Donc l'hypothèse d'hétéroscédasticité n'est pas vérifiée.

3- Multi colinéarité :

En regardant la matrice de corrélation, on soupçonne une multi colinéarité surtout entre TAX et RAD.

Pour corriger la multi colinéarité, on fait la régression sur toutes les variables explicatives et on supprime la variable ayant le plus grand VIF s'il est supérieur à 5.

Tableau 1 : VIF de la première régression avec toutes les variables explicatives

VARIABLES	VIF
CRIM	1.792192
ZN	2.298758
INDUS	3.991596

CHAS	1.073995
NOX	4.393720
RM	1.933744
AGE	3.100826
DIS	3.955945
RAD	7.484496
TAX	9.008554
PTRATIO	1.799084
B	1.348521
LSTAT	2.941491

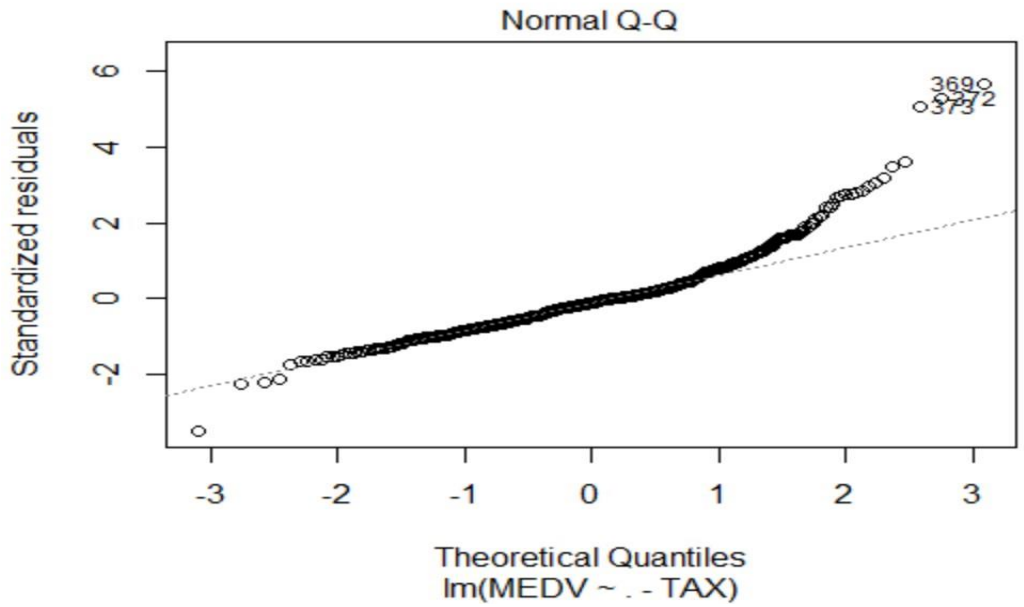
La variable TAX a un VIF qui est nettement supérieur à 5, on la supprime. Tableau 2 : VIF des variables après suppression de TAX

VARIABLES	**VIF**
CRIM	1.791940
ZN	2.184240
INDUS	3.226015
CHAS	1.058220
NOX	4.369271
RM	1.923075
AGE	3.098044
DIS	3.954446
RAD	2.837494
PTRATIO	1.788839
B	1.347564
LSTAT	2.940800

Maintenant tous les VIF sont inférieurs à 5. Le problème de multi colinéarité est fortement réduit.

4- Validation du modèle

Figure 9 : Normalité des résidus



La droite des quantiles de la loi normale colle bien avec les résidus studentisés. Et pas plus de 5% d'entre eux ne sortent dans l'intervalle [-2, 2]. Donc il semblerait que les résidus suivent une loi normale.

On va vérifier cela avec le test de Kolmogorov-Smirnov. On obtient les résultats ci-dessous :

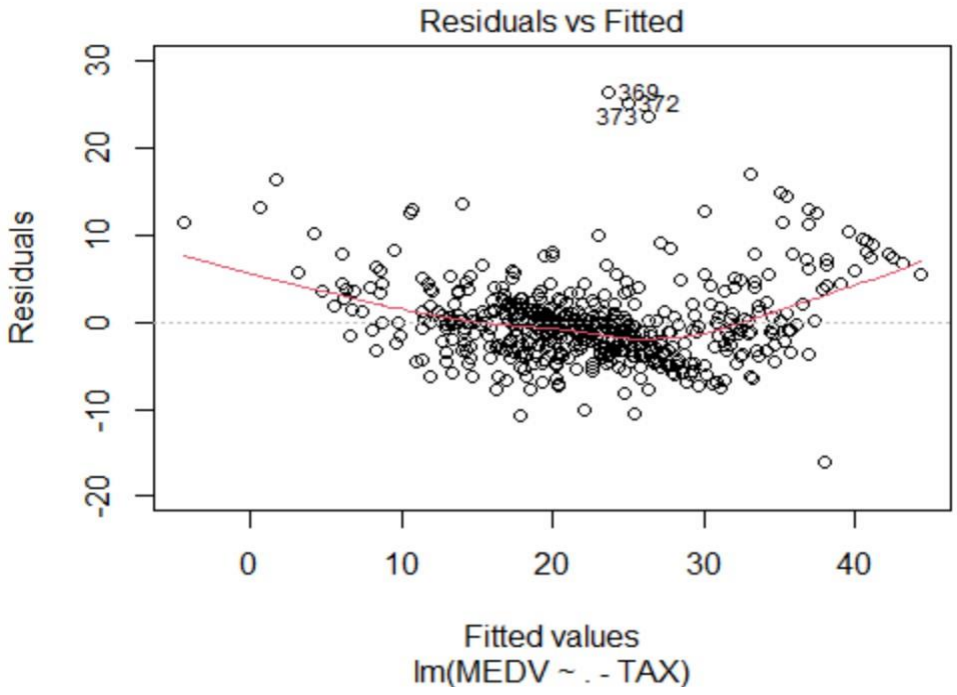
Tableau 3 : Test de Kolmogorov-Smirnov (KS test)

Statistique du test	P-value	Hypothèse nulle
0.32829	< 2.2e-16	Normalité

Le test de KS conclut la non-normalité des résidus. Mais compte tenu de la taille de notre échantillon on peut supposer la normalité des résidus.

Homoscédasticité :

Figure 10 : Graphique des résidus par rapport aux valeurs prédites



A partir de ce graphique croisant les valeurs prédites avec les résidus studentisés, on voit bien que la variance n'est pas constante pour tous les individus.

Le test de Breusch-Pagan va nous permettre de le confirmer. Ci-dessous, on a les résultats de ce test :

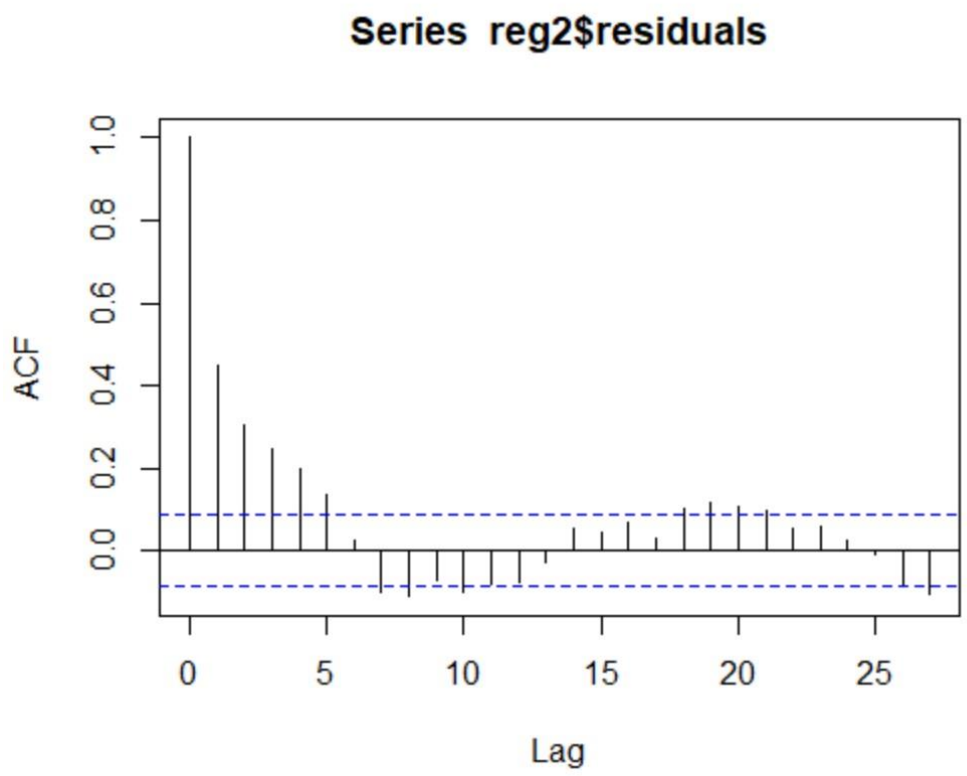
Tableau 4 : Test de Breusch-Pagan (BP test)

Statistique du test	Degré de liberté	P-value	Hypothèse nulle
62.254	12	8.736e-09	Homoscédasticité

Le test de Breusch-Pagan vient nous confirmer cette hétéroscédasticité avec une p-value très proche de 0.

Autocorrélation des résidus

Figure 11 : ACF des résidus



Il y' a des barres qui sortent de l'intervalle de confiance donc les résidus sont autocorrélés.

Les deux tests suivants vont permettre de le confirmer.

Tableau 5 : Test de Breusch-Godfrey (BG test)

Statistique du test	Degré de liberté	P-value	Hypothèse nulle
114.5	2	< 2.2e-16	Non autocorrélation

Tableau 6 : Test de Box-Pierce (BG test)

Statistique du test	Degré de liberté	P-value	Hypothèse nulle
101.87	1	< 2.2e-16	Indépendance

Les p-values des deux tests sont significatives donc rejettent les hypothèses nulles. Et ainsi on conclut l'autocorrélation des résidus.

Correction de l'hétéroscédasticité et de de l'autocorrélation

Pour la correction, on va remplacer la covariance initiale par celle de White, qui est robuste à l'hétéroscédasticité et à l'autocorrélation des résidus. Ainsi avec cette dernière on peut faire des inférences valides.

On obtient les résultats suivants :

Figure 12 : Estimation avec la matrice de White

```
z test of coefficients:

              Estimate  Std. Error z value  Pr(>|z|)
(Intercept)  34.6286431   8.4772784  4.0849 4.410e-05 ***
CRIM         -0.1067339   0.0349990 -3.0496 0.0022912 **
ZN            0.0363707   0.0141660  2.5675 0.0102448 *
INDUS        -0.0677782   0.0524643 -1.2919 0.1963941
CHAS          3.0292314   1.3095830  2.3131 0.0207157 *
NOX         -18.7012125   3.9840548 -4.6940 2.679e-06 ***
RM            3.9116902   0.8984615  4.3538 1.338e-05 ***
AGE          -0.0006054   0.0173445 -0.0349 0.9721558
DIS          -1.4883027   0.2265618 -6.5691 5.063e-11 ***
RAD           0.1345756   0.0515988  2.6081 0.0091042 **
PTRATIO      -0.9851286   0.1207233 -8.1602 3.344e-16 ***
B             0.0095464   0.0028089  3.3987 0.0006772 ***
LSTAT        -0.5222095   0.1054843 -4.9506 7.399e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

On voit bien qu'en corrigeant l'hétéroscédasticité et l'autocorrélation, les valeurs des coefficients restent les mêmes. Ce qui confirme la théorie.

La matrice de variance covariance initiale a été remplacée par celle de White, qui est robuste à l'hétéroscédasticité et à l'autocorrélation des résidus. Ainsi avec cette dernière on peut faire des inférences valides.

On voit que même avec correction des problèmes, les variables 'INDUS' et 'AGE' restent non significatives.

Les variables qui expliquent la valeur médiane des maisons par ville : CRIM, ZN, CHAS, NOX, RM, DIS, RAD, PTRATIO, B, LSTAT.

- Le signe **négatif** de la variable **CRIM** était attendu. Plus le taux de criminalité est élevé dans une ville, moins les gens ont envie de s'y installer et cela influence négativement le prix des maisons dans cette ville (car l'offre sera nettement supérieure à la demande).
- La variable "rivière Charles" (**CHAS**) a un signe **positif**. Cela pourrait s'expliquer par le fait que les villes bordées par la rivière soient à des prix élevées dû à leurs attractivités (belle vue, lieu de promenade, petite partie de pêche…), ce qui rend la valeur médiane plus élevée.
- Le signe de la variable **NOX** qui représente la concentration d'oxydes nitriques est **négatif**, on pourrait avancer l'hypothèse selon laquelle les villes polluées ne sont pas très désirées par les gens pour des raisons de santé.
- La variable **nombre moyen de pièces** a un signe **positif**. On pouvait s'y attendre car en général plus une maison a beaucoup de pièces plus son prix est conséquent.
- La variable **DIS** (distances pondérées vers cinq centres d'emplois de Boston) a un coefficient **négatif,** pas surprenant. Car plus cette distance est élevée plus la ville manque d'opportunités professionnelles. Et les gens ont plus tendance à habiter les lieux opportuns.
- Le signe de la variable ratio élèves-enseignant par ville (PTRATIO) est **négatif**. Cette variable mesure la qualité de l'enseignement de la ville. Cette variable reste contre intuitif.
- Plus la proportion de noirs augmente, plus la variable B augmente. Cette dernière est de signe positif et cela pourrait dire que plus il de noirs, plus le prix augmente. Connaissant les états-Unis, ceci reste contre intuitif.
- La variable pourcentage de statut inférieur de la population (**LSTAT**) a un signe **négatif**.

Compte tenu des coefficients estimés et de leurs p-value, on peut dire que la concentration d'oxydes nitriques, le nombre moyen de pièces par maison, les distances vers 5 centres d'emplois de Boston sont les premiers facteurs qui influencent la valeur médiane des maisons par ville. Il faut notamment souligner que pour les villes bordées par la rivière Charles, la valeur médiane des maisons augmente de 2960 dollars par rapport à celle des autres villes.

Concernant la variable ratio enseignants-élèves, elle est corrélée significativement dans le sens négatif. Le fait que le coefficient qui lui est assigné soit petit par rapport aux autres

en valeur absolu indique que la qualité de l'enseignement par ville n'est pas la priorité majeure pour les acheteurs.

Enfin pour faciliter l'utilisation de notre modèle par des novices nous avons décider de le déployer sur un interface graphique html. Le déploiement a été fait à l'aide de R Shiny et l'application est accessible à partir du lien : https://film.shinyapps.io/PredictionPrix/.